DOCUMENT RESUME

ED 400 298 TM 025 615

AUTHOR Johanson, George A.; Rich, Charles E.
TITLE Grading Large Classes: An Application of Linear Equating to Percentage-Correct Grading Decisions.
PUB DATE Apr 91
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Class Size; College Students; Criterion Referenced Tests; *Difficulty Level; *Equated Scores; Grades (Scholastic); *Grading; Higher Education; Scoring; *Standards
IDENTIFIERS Absolute Values; *Anchor Tests; Large Scale Assessment; *Linear Equating Method; Number Right Scoring; Standard Setting

ABSTRACT

Assigning letter grades in a consistent manner to tests in large classes across semesters is problematic if absolute grading standards are used. It may be unreasonable to implement the usual standard-setting approaches recommended for large-scale criterion-referenced testing due to both time constraints and a desire to have criteria that appear uniform. However, percentage-correct grading standards cannot be fairly applied without adjustment to tests of differing difficulty. The suggestion is made that linear equating with an anchor test design may be an appropriate procedure for making the adjustment in many such circumstances. An example using real data from final examinations of an introductory social science course taken by 597 students in the winter and 609 students in the spring is examined. Apparently small differences in test difficulty are seen to yield large differences in the grades assigned when scores are put on a common scale. (Contains 2 tables and 10 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
GEORGE JOHANSON

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 298

Grading Large Classes: An Application of Linear Equating to Percentage-Correct Grading Decisions

A paper presented at the 1991 annual meeting of the National Council on Measurement in Education, Chicago, IL

George A. Johanson & Charles E. Rich
Ohio University

TM025615

Abstract

Assigning letter grades in a consistent manner to tests in large classes across semesters is problematic if absolute grading standards are used. It may be unreasonable to implement the usual standard setting approaches recommended for large-scale criterion-referenced testing due to both time constraints and a desire to have criteria that appear uniform. However, percentage-correct grading standards cannot be fairly applied without adjustment to tests of differing difficulty. The suggestion is made that linear equating with an anchor-test design may be an appropriate procedure for making the adjustment in many such circumstances. An example using real data is examined; apparently small differences in test difficulty are seen to yield large differences in the grades assigned when scores are put on a common scale.

Grading Large Classes: An Application of Linear Equating to Percentage-Correct Grading Decisions

Objectives

A relative or norm-referenced (NR) approach to grading is sometimes recommended (Ebel, 1979; Thorndike & Hagen, 1961); there are also calls for the use of absolute standards or criterion-referenced (CR) approaches (Hadley & Vitale, 1985; Kubiszyn & Borich, 1990). If the decision is made to use CR grading, then standards must be established. It would make sense to have possibly different standards for each test and to use one or more of the recommended methods available to set the criteria (Mills & Melican, 1988; Livingston & Zieky, 1989). However, many teachers and institutions seem to prefer, or are at least more familiar with, percentage-correct standards. Regardless of the grading system, it is necessary to make every effort to ensure that the grading is both fair and reliable.

It is often neither possible nor desirable to use identical tests each time a course is offered for

reasons of test security, evolving curricula, and instructional differences. Nevertheless, it is often the case that a subset of the items are the same, or can be made the same, as in tests for students in previous courses. The common items make it possible to use one group of students as a norming group and to put the scores of more recent groups of students on a common scale with this previous group. Differences in the difficulty levels of the two tests and in the achievement of the two groups are adjusted by the equating. Such a method of grading is a compromise between purely NR and CR techniques and is based on methods commonly used in large-scale achievement testing where, for instance, several forms of a test must be put on a common scale to permit comparisons between students taking these different forms.

An Example with Real Data

Data was obtained from both Winter, 1989 and Spring, 1990 final examinations of an introductory social science course (multiple-sections) at a midwestern university. Each examination had 75 four-option multiple-choice items. There were 23 common items and 52 unique items on the tests. The

Winter course had 597 students take the final examination and the Spring course had 609 students take a different (but for the common 23 items) 75-item examination. The tests were machine-scored and a common-item Tucker equating was performed (Kolen & Brennan, 1987) using the micro-computer software LEQUATE (Waldron, 1988) with an internal anchor-test design. The Spring examinations were put on the scale of the Winter examinations, both graded using percentage-correct criteria. The Winter examination was judged to be a suitable norming group since the test difficulty and percentage-correct grading standards resulted in an acceptable distribution of letter grades for this course.

Results

Both the Winter and Spring terms used two forms (A, B) of a final examination with identical items in different orders to reduce cheating. The Winter examination forms were alternately distributed to the students; differences between the mean scores of the two forms were non-significant (μ_A=58.40, μ_B=57.50, t=1.86, df=595, p=0.063). Similar results were seen in the Spring with two differently-ordered forms

(μ_A=58.94, μ_B=59.20, t=-0.51, df=607, p=0.611). No equating was deemed necessary across forms A and B of either test, so the data were pooled within both the Winter and Spring courses. A recent paper by Dorans & Lawrence (1990) suggests a method of determining whether an equating under these circumstances is warranted. The procedure was implemented with this data and confirmed the decision that no equating was necessary between forms for either Winter or Spring.

The difference between the mean scores of the Winter and Spring examinations (μ_w=57.96, μ_s=59.07) was statistically significant (t=-3.14, df=1204, p =0.002), though only about one point. The mean scores on the 23 common items (15.90 and 15.83, respectively) indicate that the two groups of students may have had similar levels of achievement and that the unique items on the Spring test may have been slightly easier than the unique items on the Winter test.

The reliabilities (KR-20) for the two Winter forms were both 0.721; for the two Spring forms, the values were 0.742 and 0.762. Grades were calculated for the Spring class using both equated and unequated scores using the following fixed percentage-correct grading

categories of:

A= 93-100% A-=90-92% B+=87-89% B =83-86%

B-=80-82% C+=77-79% C =73-76% C-=70-72%

D+=67-69% D =63-66% D-=60-62% F =0-59%.

Since the Spring examination was approximately one point easier than the Winter examination, equated Spring scores were sometimes lower than the unequated Spring scores (Table 1). The slope of the equating

---

insert Table 1 about here

---

line was 0.934 and the intercept was 2.691. The equating used a synthetic population with equal weights (0.5, 0.5) for the Spring and Winter (Kolen & Brennan, 1987). A similar equating resulted from using weights of 0.0 and 1.0 (slope=0.934, intercept=2.700). When the grading standards were applied to both the equated Spring scores and the unequated Spring scores, 288 (47.29%) out of the 609 unequated grades were lowered one grading category using equated scores (Table 2).

---

insert Table 2 about here

If mean letter grades are calculated (using the scale: F=0, D-=1, D=2,..., A-=10, A=11) then the mean Winter grade was a C+ (6.00) while the mean unequated Spring letter grade was B-/C+ (6.51). The mean equated Spring grade, however, was the same C+ (6.01) as in the Winter.

Conclusion and Significance

Since the mean unequated scores of the students or, equivalently, the mean difficulties of the items were somewhat similar from Winter to Spring, it was surprising that the grades of so many students (47.29%) would be affected. Certainly the number and closeness of the grading categories was a factor. Nevertheless, if the data we present is rather typical, and we have no reason to believe otherwise, then it would be wise to use scaled scores for grading decisions to allow only intentioned differences in test difficulty to affect grading decisions.

An additional advantage of this method of grading is the ability to detect changes in student achievement over time. Since even 'absolute' grades tend to be relative in the sense that similar grading distributions are seen at institutions with widely

differing student admissions policies (Aiken,1972), it is likely that faculty adjust their standards to the ability level of their students. While such adjustments may well be desirable, when they are made unconsciously it is impossible to detect how achievement is impacted by changes in admissions policies, varying attention to prerequisites, the effect of remediation programs, the use of graduate assistants, text and/or curriculum changes, and so on. If scores on examinations are equated or scaled to a reference group, then differences in achievement over time may be observed.

A final advantage of this method of grading is seen when absolute standards are used and a particular test proves to be unusually, perhaps unacceptably, easy or difficult. With an equating methodology, it is possible to avoid the difficult decision to either use an arbitrary adjustment or to give a disproportionate number of high or low grades.

References

Aiken, L. R. (1972). The grading behavior of a college faculty. In V. H. Noll, D. P. Scannell, & R. P. Noll (Eds.), Introductory readings in educational measurement. Boston: Houghton-Mifflin.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. Applied Measurement in Education, 3(3), 245-254.

Ebel, R. L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hadley, M., & Vitale, P. (1985). Evaluating student achievement. (ERIC Document Reproduction Service No. ED 285 878)

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied Psychological Measurement, 11, 263-277.

Kubiszyn, T., & Borich, G. (1990). Educational testing and measurement (3rd ed.). Glenview IL: Scott, Foresman.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2(2), 121-141.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. Applied Measurement in Education, 1(3), 261-275.

Thorndike, R. L., & Hagen, E. (1961). Measurement and evaluation in psychology and education. New York: Wiley.

Waldron, W. J. (1988). LEQUATE: Linear equating for the common-item non-equivalent-populations design. Applied Psychological Measurement, 12, 323.

Table 1
Equating Table for Spring Scores to the Winter Scale

Raw Score	Equated Score	Raw Score	Equated Score
00	02.69	38	38.17
01	03.63	39	39.11
02	04.56	40	40.04
03	05.49	41	40.97
04	06.43	42	41.91
05	07.36	43	42.84
06	08.29	44	43.77
07	09.23	45	44.71
08	10.16	46	45.64
09	11.09	47	46.58
10	12.03	48	47.51
11	12.96	49	48.44
12	13.90	50	49.38
13	14.83	51	50.31
14	15.76	52	51.24
15	16.70	53	52.18
16	17.63	54	53.11
17	18.56	55	54.05
18	19.50	56	54.98
19	20.43	57	55.91
20	21.37	58	56.85
21	22.30	59	57.78
22	23.23	60	58.71
23	24.17	61	59.65
24	25.10	62	60.58
25	26.03	63	61.52
26	26.97	64	62.45
27	27.90	65	63.38
28	28.84	66	64.32
29	29.77	67	65.25
30	30.70	68	66.18
31	31.64	69	67.12
32	32.57	70	68.05
33	33.50	71	68.99
34	34.44	72	69.92
35	35.37	73	70.85
36	36.31	74	71.79
37	37.24	75	72.72

Table 2
Equated versus Unequated Grades for Spring

Equated Grades	Unequated Grades											
	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
A	1	0	0	0	0	0	0	0	0	0	0	0
A-	3	0	0	0	0	0	0	0	0	0	0	0
B+	0	24	20	0	0	0	0	0	0	0	0	0
B	0	0	60	42	0	0	0	0	0	0	0	0
B-	0	0	0	89	38	0	0	0	0	0	0	0
C+	0	0	0	0	41	39	0	0	0	0	0	0
C	0	0	0	0	0	28	72	0	0	0	0	0
C-	0	0	0	0	0	0	30	50	0	0	0	0
D+	0	0	0	0	0	0	0	0	29	0	0	0
D	0	0	0	0	0	0	0	0	5	12	0	0
D-	0	0	0	0	0	0	0	0	0	8	5	0
F	0	0	0	0	0	0	0	0	0	0	0	13